Online Unsupervised Classification with Model Comparison in the Variational Bayes Framework for Voice Activity Detection

AUTHOR(S):

Cournapeau, David; Watanabe, Shinji; Nakamura, Atsushi; Kawahara, Tatsuya

CITATION:

Cournapeau, David ...[et al]. Online Unsupervised Classification with Model Comparison in the Variational Bayes Framework for Voice Activity Detection. IEEE Journal of Selected Topics in Signal Processing 2010, 4(6): 1071-1083

ISSUE DATE:

2010-12

URL:

http://hdl.handle.net/2433/131746

Online Unsupervised Classification With Model Comparison in the Variational Bayes Framework for Voice Activity Detection

David Cournapeau, Shinji Watanabe, *Member, IEEE*, Atsushi Nakamura, and Tatsuya Kawahara, *Senior Member, IEEE*

***Abstract*—A new online, unsupervised method for Voice Activity Detection (VAD) is proposed. The conventional VAD methods often rely on heuristics to adapt the decision threshold to the estimated SNR. The proposed VAD method is based on the Variational Bayes (VB) approach to the online Expectation Maximization (EM), so that it can automatically adapt the decision level and the statistical model at the same time. We consider two parallel classifiers, one for the noise-only case, and the other for speech-and-noise case. Both models are trained concurrently and online using the VB framework. The VB framework also provides an explicit approximation of the log evidence called free energy. It is used to assess the reliability of the classifier in an online fashion, and to decide which model is more appropriate at a given time frame. Experimental evaluations were conducted on the CENSREC-1-C database designed for VAD evaluations. With the effect of the model comparison, the proposed scheme outperforms the conventional VAD algorithms, especially in the remote recording condition. It is also shown to be more robust with respect to changes of the noise type.**

***Index Terms*—, Sequential estimation, speech analysis, variational Bayes (VB), voice activity detection (VAD).**

I. Introduction

VOICE activity detection (VAD) is a task of automatically segmenting speech boundaries from audio signals, and is important for many speech applications, such as speech coding and automatic speech recognition (ASR) [1]. In noisy environments, it is often observed that the number of insertion errors in the ASR system increases because of false detection of noise segments as speech [2], thus a noise-robust VAD becomes crucial for overall performance of ASR systems.

VAD has been recently tackled in a supervised training context, where labeled training data are assumed to be available for training classifiers such as Gaussian mixture models (GMMs) [3], linked hidden Markov models (HMMs) [4] and support vector machines (SVMs) [5]. However, assuming training data with labels is not always practical, for example, when the acoustic condition of the target environment is unknown or changes in time. If the training data and testing data have significantly different statistical characteristics, degradation of performance is generally observed. In this paper, we investigate an approach of unsupervised, online classification, where no training data is assumed to be available. Such classifiers often rely on a threshold for the frame-level classification [6]. This threshold is adapted to the noise level, which needs to be estimated separately. As noted in [7], this adaptation depends on heuristics on the noise floor estimation. The scheme of unsupervised classification based on online expectation–maximization (EM) provides a solid framework to the problem because it can automatically adapt the decision level to the signal [8]. However, the underlying statistical model is not flexible in a sense that it assumes that speech and noise data are constantly available. The goal of this study is to develop a more flexible statistical scheme which can adaptively switch the model for the case where only noise data are available. For this purpose, we introduce a reliability measure derived from statistical model comparison.

In the proposed scheme, we assume a scalar feature is used. Specifically in this paper, we adopt high-order statistics (HOS [6], [9]) which will be explained in Section IV. Both speech and noise distributions are assumed to follow a normal distribution whose mean and variance change in time. In the case where both speech and noise are present in the signal, the model is then equivalent to a binary mixture of Gaussians, whose parameters are estimated in an online fashion. On the other hand, when only noise is present in the signal, the problem is reduced to online estimation of a normally distributed random variable. An example of the distributions of the feature dealt within this framework is shown in Fig. 1. Hence, an online model comparison is defined to distinguish between those two situations for VAD, and we propose to use the Variational Bayes (VB) framework to solve the problem. The VB framework [10] provides an explicit approximation of the log-evidence called free energy, which can be used for model comparison [11], [12]. Online extension of VB based on a stochastic approximation [13] of the free energy [14] is used for online model comparison, to take into account possible changes in the acoustical environment. As the VB framework also provides an explicit form of the posterior of model parameters, both parameter estimation and model comparison are conducted with the same underlying statistical scheme. The proposed VAD method is thus based on estimation of two models concurrently (one for the speech-and-noise situation, and the other for the noise-only

Manuscript received December 03, 2009; revised January 19, 2010; accepted February 08, 2010. Date of publication September 27, 2010; date of current version November 17, 2010. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Samy Bengio.

D. Cournapeau is with the School of Informatics, Kyoto University, Kyoto 606-8501, Japan, and also with the NTT Communication Science Laboratories, NTT Corporation, Kyoto 619-0237, Japan (e-mail: cournape@gmail.com).

S. Watanabe and A. Nakamura are with the NTT Communication Science Laboratories, NTT Corporation, Kyoto 619-0237, Japan (e-mail: watanabe@cslab.kecl.ntt.co.jp; ats@cslab.kecl.ntt.co.jp).

T. Kawahara is with the School of Informatics, Kyoto University, Kyoto 606-8501, Japan (e-mail: kawahara@ar.media.kyoto-u.ac.jp).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2010.2080821

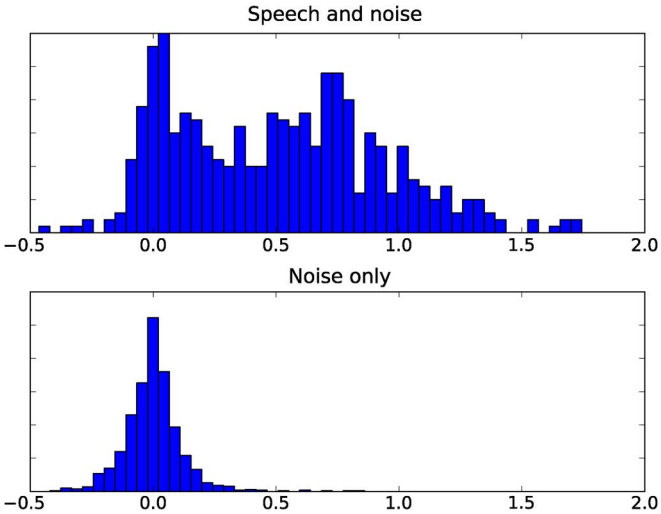

Fig. 1. Histogram of the feature used in this work: histogram for the speech-plus-noise parts (top) and noise-only (bottom). The exact feature is described in Section IV, and the dataset in Section V.

situation), which are assessed with the online free energy and selectively used, for every time frame.

The organization of the paper is as follows. Online EM for unsupervised classification as well as its limitations for VAD is first presented in Section II. We review the VB-EM framework for explicit optimization of the free energy for model comparison in Section III. Its application to the VAD task is described in Section IV, and the proposed VAD method is evaluated on the CENSREC-1-C database in Section V, where it is compared to conventional VAD methods and the online EM classifier without the model comparison.

II. Online EM for Unsupervised Classification

When we try unsupervised classification without training data, the classification often relies on a threshold adapted from the noise floor. The threshold is estimated and updated from the background noise level, and the frame-level speech/non-speech classification is converted to speech boundaries using a hang-over scheme. This is the most straightforward method for unsupervised classification. As noted in [7], the noise floor is often estimated using heuristics.

When we use a statistical framework instead, presence/absence of speech can be regarded as the realization of a binary random variable h, and the feature values as the realization of a random variable (or vector for multidimensional features) x. If we assume each class to be normally distributed, the observation model is a GMM, and estimation can be tackled using the EM algorithm [15] applied to latent models. As each iteration of the EM algorithm requires the whole dataset, it cannot be used for online classification where the observations come one after the other. An online extension of the EM algorithm based on a stochastic approximation has been proposed recently [16], [17], and we applied it to VAD in [8]. In this section, we will briefly review the principles of this online extension, as well as its limitations when applied to VAD.

A. EM Algorithm

The maximum-likelihood estimation (MLE) is hard to compute explicitly for latent models, and the EM algorithm is a method to optimize the log-likelihood for a wide range of models where the MLE is intractable. Given N independent and identically distributed (i.i.d.) observations $x \triangleq x_1, \ldots, x_N$, the log-likelihood L of θ is defined as

$$L(\theta) \triangleq \ln p(x;\theta) = \sum_{n=1}^{N} \ln p(x_n;\theta). \tag{1}$$

The key principle of EM applied to the MLE framework is to build an auxiliary function $Q(\theta)$ which is easier to maximize than the observed log-likelihood $L(\theta)$, while its maximization will give a reasonable estimate of the MLE applied directly to L. The standard EM algorithm defines the function Q as the expected log-likelihood of some complete data (x,h) conditionally on the observations x only, where $h \triangleq (h_1,\ldots,h_N)$ are latent variables

$$Q_{\theta^i}(\theta) \triangleq \mathbb{E}\left[\ln p(x,h;\theta)|x;\theta^i\right] \tag{2}$$

$$\theta^{i+1} \triangleq \arg\max_{\theta} Q_{\theta^i}(\theta) \tag{3}$$

where θ^i is the parameter estimated at the i^{th} iteration. Iteratively running (2) and (3) gives a sequence $\{\theta^i\}$ which converges to a local maximum of the log-likelihood L in general settings [18]. In particular, if the complete data (x,h) follow a density in the (Natural) Exponential Family[1] (EF, [19])

$$p(x;\theta) \triangleq \int p(x,h;\theta)dh \tag{4}$$

$$\ln p(x,h;\theta) \triangleq \langle s(x,h),\theta\rangle + s_0(x,h) - \psi(\theta) \tag{5}$$

where s is a function of x of the same dimension as θ and is sufficient statistic (SS) for θ, $\langle .,. \rangle$ the scalar product, ψ a function of θ, and s_0 another function of x, then the computation of Q is reduced to the conditional-expectation of $s(x,h)$ under the density $p(h|x)$:

$$Q_{\theta^i}(\theta) = \mathbb{E}\left[\langle s(x,h),\theta\rangle + s_0(x,h) - \psi(\theta)|x;\theta^i\right] \tag{6}$$

$$\propto \left\langle \overline{s}(x;\theta^i),\theta\right\rangle - \psi(\theta) \tag{7}$$

where the terms which do not depend on θ have been omitted, and $\overline{s}(x;\theta^i)$ is defined as the expected (or averaged) SS under the parameter θ^i:

$$\overline{s}(x;\theta^i) \triangleq \frac{1}{N}\sum_{n=1}^{N} \overline{s}(x_n;\theta^i) \tag{8}$$

$$\overline{s}(x_n;\theta^i) \triangleq \mathbb{E}\left[s(x_n,h_n)|x_n;\theta^i\right] \tag{9}$$

Noting f the function

$$f(s) \triangleq \arg\max_{\theta} \left[\langle s,\theta\rangle - \psi(\theta)\right] \tag{10}$$

The EM algorithm [(2) and (3)] can then be succinctly written as follows:

$$\theta^{i+1} = f\left(\overline{s}(x;\theta^i)\right) \tag{11}$$

Practical implementations of the EM algorithm are possible when Q can be optimized efficiently and explicitly, which is the case for GMM and HMM with Gaussian mixture distributions.

[1] x is also said to follow a density in the Exponential Hidden Family (EHF).

B. Online EM

When the observation comes one after another, and the classification is needed after each observation, the EM algorithm cannot be used as it is because each iteration of the E step (2) needs all the data at once. Online extensions of the EM algorithm have been studied, first to alleviate the relatively intensive computational and memory cost, and later for online estimation problems. A recent approach is based on recursively approximating Q itself, while keeping the M step essentially the same [16], [17]. The online approximation $\hat{Q}_n$ of Q is based on the following recursion:

$$\hat{Q}_{n+1}(\theta) = \hat{Q}_n(\theta) + \gamma_{n+1}\left[\mathbb{E}\left[\ln p(x_{n+1}, h_{n+1}; \theta)|x_{n+1}; \theta_n\right] - \hat{Q}_n(\theta)\right] \quad (12)$$

where γ_n is a learning parameter. The M step is kept the same as for the offline EM, that is $\hat{\theta}_{n+1}$ is set as the maximum of $\hat{Q}_{n+1}$; each iteration of this procedure is repeated once for each new observation x_n (the iteration index and the sample index are now the same). When the complete data follow a density in the EF, the online update (12) can then be reduced to online estimation of the averaged SS $\hat{s}_n$

$$\hat{s}_{n+1} = \hat{s}_n + \gamma_{n+1}\left(\overline{s}(x_n; \hat{\theta}_n) - \hat{s}_n\right) \quad (13)$$

$$\hat{\theta}_{n+1} \triangleq f(\hat{s}_{n+1}). \quad (14)$$

This method can be considered as a stochastic approximation of the expected log-likelihood [16], [17], [20]. The learning parameter γ_n must follow the usual conditions for the stochastic approximation to converge [13]

$$\gamma_n > 0 \quad (15)$$

$$\sum_n \gamma_n = \infty \quad (16)$$

$$\sum_n \gamma_n^2 < \infty \quad (17)$$

The properties of the procedure defined by (13) and (14), including theoretical considerations on convergence can be found in [20]. In particular, it is proved that the series $\{\hat{\theta}_n\}$ defined by (13) and (14) converges to a stationary point of the Kullback–Leibler divergence between the observation density and the model density under some technical assumptions. Since $\overline{s}(x_n; \hat{\theta}_n)$ only depends on the observation at time n, this procedure defines a practical online estimation when both $\overline{s}$ and f can be computed explicitly and efficiently. In this case, the practical implementation can be derived from the conventional, batch EM implementation [20].

In the case of a Gaussian mixture, in which $w_{n,k}$, $\mu_{n,k}$ and $\Omega_{n,k}$ represent the component weight, mean and precision for component k, the online-EM procedure becomes

$$\hat{s}_{n+1,k} = \begin{pmatrix} \hat{s}_{n+1,k,1} \\ \hat{s}_{n+1,k,2} \\ \hat{s}_{n+1,k,3} \end{pmatrix} = \hat{s}_{n,k} + \gamma_{n+1}\left(\begin{pmatrix} \bar{w}_k(x_{n+1}; \hat{\theta}_n) \\ \bar{w}_k(x_{n+1}; \hat{\theta}_n) x_{n+1} \\ \bar{w}_k(x_{n+1}; \hat{\theta}_n) x_{n+1} x'_{n+1} \end{pmatrix} - \hat{s}_{n,k}\right) \quad (18)$$

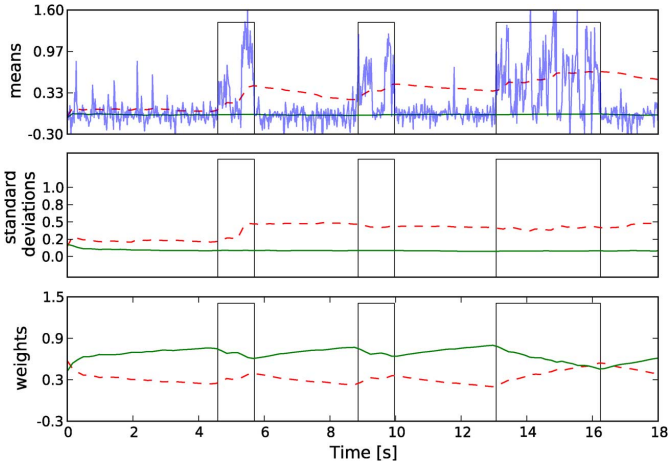

Fig. 2. Example of online-EM applied to a one dimension feature extracted from input signals. Means (top), standard deviations (middle) and weights (bottom) estimated with the online EM are displayed. The dashed line corresponds to speech, and the solid line to noise.

where $\bar{w}_k(x_{n+1}; \hat{\theta}_n)$ is the responsibility for sample n

$$\overline{w}_k(x_n; \theta^i) \triangleq p(h_n = k|x_n; \theta^i) \quad (19)$$

and their functional form is exactly the same as in the conventional EM, thus the parameters are updated as follows:

$$\forall k \in 1, \ldots, K, \quad \begin{cases} \hat{w}_{n+1,k} = \hat{s}_{n+1,k,1} \\ \hat{\mu}_{n+1,k} = \frac{\hat{s}_{n+1,k,2}}{\hat{s}_{n+1,k,1}} \\ \hat{\Omega}_{n+1,k} = \frac{\hat{s}_{n+1,k,3}}{\hat{s}_{n+1,k,1}} - \hat{\mu}_{n+1,k}\hat{\mu}'_{n+1,k} \end{cases} \quad (20)$$

The exact derivation of (18) and (20) is reviewed in Appendix A.

C. Application to Voice Activity Detection and Limitations

Online EM is used for concurrent noise/speech level estimation as follows. Each class (speech and noise) is assumed to be normally distributed, and the observed model is thus a binary mixture of Gaussian distributions, where one class $(h_n = 0)$ corresponds to noise and the other $(h_n = 1)$ to speech. Online EM gives a new set of estimated parameters for each frame, and as such the decision level d is defined as

$$d \triangleq p(h_n = 1|x_n; \hat{\theta}_n) - p(h_n = 0|x_n; \hat{\theta}_n) \quad (21)$$

which is automatically adapted on a per-frame basis. Although the assumption of Gaussianity for each class is simple, the scheme allows for the actual distribution of each class to be more complex over a long time period, since its parameters are allowed to change over time. This is shown in Fig. 2, where the online EM procedure is applied on a one-dimensional feature, which is described in Section IV-A.

This method was shown to give reasonable results in [8], but suffers from some inherent limitations. First, at the beginning of the signal, because there is only noise or speech, the training of the Bayesian classifier $\arg\max_k p(k|x_n; \hat{\theta}_n)$ is unreliable. This can be observed in Fig. 2. For the first few seconds, because no speech frame is available, the means of the binary mixture are close to each other. Once some speech frames were input to the online EM, the corresponding classifier can be used effectively (after five seconds in Fig. 2). This problem can be somewhat alleviated by using some heuristics (as used in many works, assuming

that the first seconds of the signal are noise-only), but we present a more theoretically sound solution. Moreover, when there is no speech for a long time, the means of the mixture components will become close to each other, and as such, the classifier will also be unreliable. Both problems are related to the fact that, when the Gaussian distributions of the mixture are mostly overlapping, the mixture does not properly represent two-class model as designed. In this paper, we use model comparison to detect the cases where the estimated model does not represent a two-class model.

III. Free Energy Maximization in the Variational Bayes Approach

The statistical model used in Section II can be seen as a binary mixture, whose state changes in time. To solve the problem mentioned above, we propose to use the Bayesian framework for inference, in particular for model comparison; it is used to compare whether the data are better explained by a model with one or two components in our application to VAD. This section reviews the principles of Bayesian model comparison, and the Variational Bayes (VB [10], [21]) framework to make the computation tractable.

A. Using Bayesian Inference for Model Comparison

Bayesian inference assumes parameters to be random variables, and Bayesian estimators are based on posterior probabilities.[2] Since models themselves can be seen as parameters, they can also be inferred using the observations (see for example chapter 28 in [11]). For a given Gaussian mixture model m with K components, the joint probability density function (pdf) for the observations x, the latent data h, and the parameters θ is given by the pdf $p(x,\theta,h|m)$. Bayesian estimators are then based on the posterior $p(\theta,h|x,m)$

$$p(\theta,h|x,m) = \frac{p(x,h|\theta,m)\cdot p_0(\theta|m)}{p(x|m)} \tag{22}$$

$$\propto p(x,h|\theta,m)\cdot p_0(\theta|m) \tag{23}$$

where $p_0(\theta|m)$ is a prior of the parameters given the model m. As the model m itself is also a random variable, a model posterior can be computed as well

$$p(m|x) = \frac{p(x|m)p(m)}{p(x)} \tag{24}$$

$$\propto p(x|m)\cdot p(m). \tag{25}$$

The marginalized likelihood $p(x|m)$, also called the evidence, is obtained by marginalizing over both the parameters θ and the latent variables h

$$p(x|m) = \int p(x,\theta,h|m)d\theta dh = \int p(x,h|\theta,m)\cdot p_0(\theta|m)d\theta dh \tag{26}$$

Thus, one of the advantages of Bayesian inference is that a second level of inference is possible [11], namely, once a prior on the model $p(m)$ is given, scoring different models can be done using the evidence [(26)] through (25) [11], [22], [23]. By computing the integral (26) for M different models m, these models can be compared. For example, assuming a flat prior on m $p_0(m) = 1/M$, the model posterior defined in (25) is proportional to the evidence $p(x|m)$. However, the integrals are intractable for latent models; hence, approximation schemes are needed. The VB framework provides a solution, and is based on making some assumptions on the integrand of (26). The log-evidence [(26)] is then approximated by a functional called Free Energy, which provides an explicit measure for model comparison. For a large class of models, it can be shown that the free energy and the Bayesian information criterion (BIC) converge to the same value in the limit of large samples ([24], Chapter 2). But unlike the BIC, model comparison with free energy is not based on a large number of samples hypothesis, and it can be naturally extended to the online case.

B. Free Energy for Model Comparison

The VB provides a tractable lower bound of the marginalized log-likelihood by restricting the possible functional forms of the posterior. For any function $\tilde{q}(h,\theta)$ over the hidden data h and parameters θ, the Kullback–Leibler divergence between $\tilde{q}$ and the true posterior $q \triangleq p(h,\theta|x,m)$ can be computed as follows:

$$\begin{aligned} KL(\tilde{q}||q) &\triangleq \int \tilde{q}(\theta,h)\ln\frac{\tilde{q}(\theta,h)}{p(\theta,h|x,m)}d\theta dh \\ &= \int \tilde{q}(\theta,h)\ln\frac{\tilde{q}(\theta,h)p(x|m)}{p(x,h|\theta,m)p_0(\theta|m)}d\theta dh \\ &= \ln p(x|m) - F_m[\tilde{q}] \geq 0 \end{aligned} \tag{27}$$

where the free energy $F_m[\tilde{q}]$ is a functional,[3] and is defined as

$$F_m[\tilde{q}] \triangleq \int \tilde{q}(\theta,h)\ln\frac{p(x,h|\theta,m)p_0(\theta|m)}{\tilde{q}(\theta,h)}d\theta dh \tag{28}$$

and the inequality (27) is by definition of the Kullback–Leibler divergence, and a consequence of the Jensen inequality applied to the concave function ln. Inequality (27) shows that F_m is a lower bound of the marginalized log-likelihood for any $\tilde{q}$. Thus, maximizing the negative free energy $-F_m$ with respect to the approximate distributions $\tilde{q}$ will give an approximation of the marginalized log-likelihood. As Bayesian model comparison is based on evaluating $p(x|m)$ for different models, if $-F_m$ is tight enough, it may be used in place of the marginalized log-likelihood.

C. VB-EM

The negative free energy $-F_m$ cannot be explicitly maximized as it is; the VB method is based on restricting the possible functional forms of $\tilde{q}$ to factorized forms $q(\theta,h) = q_h(h)q_\theta(\theta)$.[4] F_m is then minimized with respect to q_θ and q_h, using the tools

[2]We note $p(x|\theta)$ instead of $p(x;\theta)$; we keep the notation $p(.;a)$ when a is not considered random.

[3]We note $F[q]$ for functionals, where the argument q is a function, and $F(x)$ for functions of a variable x.

[4]Hereafter, the model parameter m will be implied and dropped from the density parameters.

of calculus of variations. Minimization of F_m is then reduced to a set of two coupled equations, similar to the EM algorithm [10]

$$q_h(h) \propto \exp\left\{\int \ln p(x,h|\theta) q_\theta(\theta) d\theta\right\} \tag{29}$$

$$q_\theta(\theta) \propto p_0(\theta) \exp\left\{\int \ln p(x,h|\theta) q_h(h) dh\right\} \tag{30}$$

As the equations are coupled, the optimization has to be done iteratively, from initial values for both q_θ and q_h. One can note that (30) is similar to the M step for EM applied to the MLE framework, except that θ is assumed random; more accurately, the right side is exactly the optimized quantity with respect to (w.r.t.) θ in the maximum *a posteriori* (MAP) context, where q_h is replaced by the usual responsibilities. Thus, the main difference compared to the simple MAP extension of EM is (29), which considers the posterior on θ as well, whereas the traditional MAP only considers a prior on θ. In the EM algorithm applied to MAP, the M step gives a point estimate $\hat{\theta}_{MAP}$ [10] (that is, MAP is not concerned with an explicit posterior over θ).

For practical computation, the VB method is usually restricted to densities within the Exponential Hidden Family (EHF), as in Section II, that is $p(x,h|\theta)$ will be given by (5). In a Bayesian context, the EHF also has the advantage to always have at least one prior conjugate to the likelihood, that is the resulting posterior has the same functional form as the prior [19]

$$\begin{aligned}\ln p_0(\theta;\tau_0,\alpha_0) &= \langle\theta,\alpha_0\rangle - \tau_0\psi(\theta) - \zeta(\tau_0,\alpha_0) \\ &\propto \langle\theta,\alpha_0\rangle - \tau_0\psi(\theta)\end{aligned} \tag{31}$$

where τ_0, α_0 are the prior's hyper-parameters. The vector α_0 has the same dimension as θ and is interpreted as a prior value on the parameter θ. The hyper-parameter τ_0 is a scalar, and can be interpreted as the pseudo count of the prior, e.g., for N observations, a ratio $\tau_0/(\tau_0+N) \ll 1$ will be representative of a weak prior. The normalization constant $\zeta(\tau_0,\alpha_0)$ can be derived by integration

$$\zeta(\tau_0,\alpha_0) = \ln \int e^{\langle\theta,\alpha_0\rangle - \tau_0\psi(\theta)} d\theta. \tag{32}$$

The main advantage of considering EHF models is the conservation of the conjugacy property; at the end of each iteration i, the posterior q_θ^{i+1} is conjugate to the prior p_0 [10]. One iteration of the VB-EM procedure applied to the EHF can be written as follows [10]:

$$q_h^{i+1}(h) = q_h(h;\overline{\theta}^{i+1}) = \prod_{n=1}^{N} q_{h_n}^{i+1}(h_n) \tag{33}$$

$$\overline{s}(x;\overline{\theta}^{i+1}) \triangleq \frac{1}{N}\sum_{n=1}^{N} \overline{s}(x_n;\overline{\theta}^{i+1}) \tag{34}$$

$$\begin{aligned}\ln q_\theta^{i+1}(\theta) &= \ln q_\theta(\theta;\tau^{i+1},\alpha^{i+1}) \\ &= \langle\theta,\alpha^{i+1}\rangle - \tau^{i+1}\psi(\theta)\end{aligned} \tag{35}$$

where we note

$$\overline{\theta}^{i+1} \triangleq \mathbb{E}_{q_\theta^i}[\theta] = \int \theta q_\theta^i(\theta) d\theta \tag{36}$$

$$q_{h_n}^{i+1}(h_n) \triangleq p(h_n|x_n,\overline{\theta}^{i+1}) \tag{37}$$

$$\begin{aligned}\overline{s}(x_n;\overline{\theta}^{i+1}) &\triangleq \mathbb{E}_{q_{h_n}^{i+1}}[s(x_n,h_n)|x_n] \\ &= \int s(x_n,h_n) q_{h_n}^{i+1}(h_n) dh_n\end{aligned} \tag{38}$$

$$\tau^{i+1} \triangleq \tau_0 + N \tag{39}$$

$$\alpha^{i+1} \triangleq \alpha_0 + N\overline{s}(x;\overline{\theta}^{i+1}). \tag{40}$$

Because q_θ itself is in the EF, $\overline{\theta}^{i+1}$ which is the first moment of θ under q_θ can be derived from the normalization factor of the posterior considered as a function of the hyper-parameters

$$\overline{\theta}^{i+1} = \bar{\theta}(\tau^i,\alpha^i) \triangleq \frac{\partial\zeta}{\partial\alpha^i}(\tau^i,\alpha^i). \tag{41}$$

A comparison of (33)–(35) with the standard equations for the EM for MLE [(8), (9) and (11)] highlights the main differences between both procedures. The functional form of the averaged sufficient statistics $\overline{s}$ is exactly the same, but it is computed for an average parameter $\overline{\theta}^{i+1}$ in the VB-EM procedure. The posterior update is replaced by hyper-parameters updates. As our main motivation for using the VB-EM framework is model comparison, the free energy also needs to be estimated. The free energy is a function of the averaged SS and both prior and posterior hyper-parameters, and thus can be estimated from the quantities computed in (38)–(40). The exact formulation in the case of a mixture of Gaussians is reviewed in Appendix B.

From an implementation point of view, the computation of $\overline{\theta}^{i+1}$ is the main additional cost of the VB-EM procedure when applied to GMM, as the hyper-parameter updates in the M step are similar to the M step of the conventional EM algorithm. For a GMM of K components of dimension d, computing $\overline{\theta}^{i+1}$ involves evaluation of $K(d+1)+1$ points of the digamma function F and the determinant of K matrices of dimension d.

A preliminary application of the VB framework to VAD was presented in [25], where we applied the free energy in a mini-batch manner to detect speech-and-noise (two Gaussians) against noise-only situations (one Gaussian). Although it gave promising improvements compared to the conventional online EM algorithm, the mini-batch application of the free energy is ad-hoc. Detecting speech-and-noise against noise-only situation on a frame-per-frame basis, using the same underlying models for both detection and classification would be more appropriate. The next section presents online extensions of the VB framework in that regard.

D. Online Extension

As in the conventional MLE, the VB-EM procedure requires all the data to be available at once. Several online extensions have been studied [10], [14]. A direct online update of the posterior as in [10] has been found ineffective to track acoustical environmental changes when applied to VAD. Our proposed method for VAD is instead based on a stochastic approximation of the free energy [14].

1) Variational Bayes as Parametrized Free Energy Optimization: The online derivation of VB-EM is similar to the online extensions of the standard EM algorithm as reviewed in Section II-B. Online EM in the MLE framework was a stochastic approximation to estimate the maximum of the expected

log-likelihood. To practically derive the online extension as a stochastic approximation, the VB-EM procedure may be expressed in a form similar to (11), that is as the optimization of a function with respect to a set of parameters. As noted in [23] and explicitly conducted in [14], the VB-EM version of (33)–(35) can be derived directly from the optimization of a parametrized free energy, noted F_m^p, with respect to the hyper-parameters (reviewed in Appendix C).

In this form, the VB-EM procedure is similar to the conventional EM algorithm, where $-F_m^p$ plays the same role as Q, and the hyper-parameters (τ, α) play the same role as θ. The relationship with the EM algorithm is clear when the VB-EM procedure [(33)–(35)] is succinctly written as

$$\overline{\theta}^{i+1} \triangleq \overline{\theta}(\tau^i, \alpha^i) \tag{42}$$

$$\overline{s}(x_n; \overline{\theta}^{i+1}) = \mathbb{E}_{q_{h_n}^{i+1}}\left[s(x_n, h_n)|x_n\right] \tag{43}$$

$$\begin{pmatrix} \tau^{i+1} \\ \alpha^{i+1} \end{pmatrix} = g\left(\overline{s}(x; \overline{\theta}^{i+1})\right) \tag{44}$$

where $g : s \mapsto g(s)$ is a function which depends only on the hyper-parameters of the prior; the functional form is entirely determined by (τ_0, α_0). Both $\overline{\theta}$ and g functional forms are kept the same between iterations of the VB-EM procedure.

As the VB-EM procedure [(42)–(44)] is equivalent to the minimization of F_m^p w.r.t. hyper-parameters, an online derivation becomes similar to online EM [(14)]: making a stochastic approximation to minimize F_m^p as a function of the hyper-parameters, and defining online hyper-parameter updates as the minimization of F_m^p at every frame, similarly to how online EM was described in Section II-B. This is the approach developed in [14].

2) Online VB-EM: The online extension of the VB method is thus in principle similar to the online extension of EM applied to the MLE. F_m^p is recursively approximated by $\widehat{F}_{mn}^p$ as Q was by $\hat{Q}_n$

$$\widehat{F}_{mn+1}^p(\tau,\alpha) = \widehat{F}_{mn}^p(\tau,\alpha) + \gamma_{n+1} \times \left[\mathbb{E}\left[\ln\left(\frac{p(x_{n+1}, h_{n+1}, \theta)}{q_h^p\left(h_n; \overline{\theta}(\tau,\alpha)\right) q_\theta^p(\theta; \tau, \alpha)}\right)\right] - \widehat{F}_{mn}^p(\tau,\alpha)\right] \tag{45}$$

and the estimated hyper-parameters $\hat{\tau}_n$ and $\hat{\alpha}_n$ are defined as the values which minimize $\widehat{F}_{mn}^p$. Thus, the online procedure for VB-EM can be written as follows:

$$\hat{\theta}_{n+1} \triangleq \overline{\theta}(\hat{\tau}_{n+1}, \hat{\alpha}_{n+1}) \tag{46}$$

$$\hat{s}_{n+1} \triangleq \hat{s}_n + \gamma_{n+1}\left[\overline{s}(x_{n+1}; \hat{\theta}_n) - \hat{s}_n\right] \tag{47}$$

$$\begin{pmatrix} \hat{\tau}_{n+1} \\ \hat{\alpha}_{n+1} \end{pmatrix} \triangleq g(\hat{s}_{n+1}). \tag{48}$$

Those online updates of hyper-parameters can be used to compute $\widehat{F}_{mn}^p$, which can be used as an online model comparison measure [14]. For online VB-EM, a practical implementation of the stochastic approximation requires an explicit relationship between the updated hyper-parameters (corresponding to the parameters in the case of online EM) and the averaged sufficient statistics, that is g as defined in (44) must be made explicit.

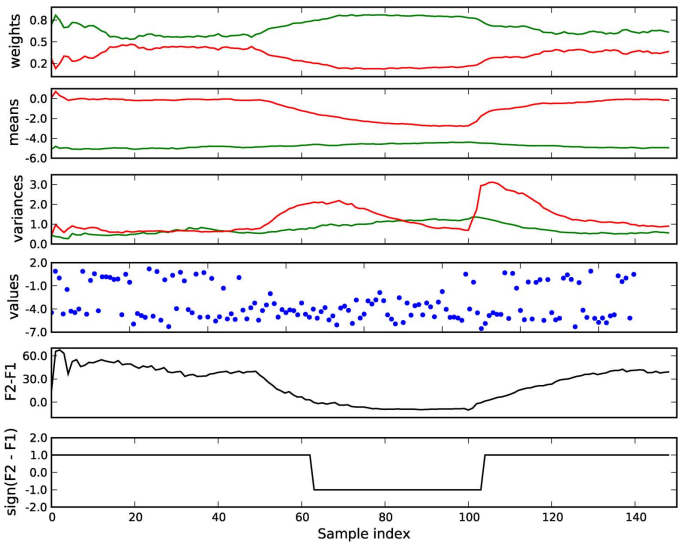

Fig. 3. Online VB-EM procedure applied on high SNR speech data. The bottom axis is the relative order between F_1 and F_2: if positive, $F_2 > F_1$, and $F_1 > F_2$ otherwise. The sampled model state is changed at sample 50 and 100.

3) Application to Mixture of Gaussians: For GMM, the stochastic approximation of the averaged SS is exactly the same as that for online EM [(65)]:

$$\hat{s}_{n+1,k} = \begin{pmatrix} \hat{s}_{n+1,k,1} \\ \hat{s}_{n+1,k,2} \\ \hat{s}_{n+1,k,3} \end{pmatrix} = \hat{s}_{n,k} + \gamma_{n+1}\left(\begin{pmatrix} \overline{s}_{k,1}(x_{n+1}; \hat{\theta}_n) \\ \overline{s}_{k,2}(x_{n+1}; \hat{\theta}_n) \\ \overline{s}_{k,3}(x_{n+1}; \hat{\theta}_n) \end{pmatrix} - \hat{s}_{n,k}\right) \tag{49}$$

with $\hat{\theta}_n$ defined by (46). The definition of g for GMM is given in Appendix B. This justifies *a posteriori* the derivation of VB-EM for the EHF, instead of directly deriving the equations for GMM from (29) and (30) in Section III, by making the relationship between $\hat{s}_n$ and the updated hyper-parameters $(\hat{\tau}_n, \hat{\alpha}_n)$ explicit. Our implementation of the VB procedure follows (42)–(44); thus, the online extension implementation [(46)–(48)] follows directly.

4) Example: An example of this procedure is shown in Fig. 3, where we sample data from an artificial binary mixture (sample displayed on the second bottom plot). The first 50 samples are sampled from a well separated mixture, then almost overlapping from sample 50 to 100, and then back to the first state starting at sample 100. The weights, means and variances are updated online. We ran VB-EM for both models with one and two components, and evaluated the online free energy in each case; the bottom axis shows value 1 when $\hat{F}_2^p > \hat{F}_1^p$ and -1 when $\hat{F}_1^p > \hat{F}_2^p$. It is observed that the online free energy can track model changes, at least on this simple example. The influence of the learning parameter is also observed, as there is a latency between the change of the sampling model (at frame 50 and 100) and the estimated online free energy. The learning

parameter γ_n is necessary to gradually forget the old input data in the online VB-EM procedure.

IV. Application to Voice Activity Detection

In this section, we describe details on the feature and the classifier used in our VAD method.

A. Enhanced Kurtosis as a Feature for VAD

As an effective scalar feature for VAD, which is modeled as a bimodal Gaussian distribution in the speech-and-noise situation, we have investigated the kurtosis and other cumulants of the linear prediction coding (LPC) residual [6], [9]. The kurtosis is defined as

$$k^X \triangleq \frac{\kappa_4^X}{\sigma^4} = \mathbb{E}\left[(X-\mu)^4\right] - 3\sigma^2 \tag{50}$$

where μ and σ^2 are the mean and variances of X. The kurtosis is robust against Gaussian noise ($k^X = 0$ for normally distributed X), and it has also been observed that using cumulants alone may not work for cases such as transient noises, as those noises are characterized by high energy over a short period of time, and the estimated kurtosis has a high value. To alleviate this problem, we have proposed in [8] to combine the kurtosis with a feature which is robust against transient noises and does not affect the desired properties of the kurtosis for speech signals and Gaussian-like noises. Specifically, we use the normalized autocorrelation, defined as below for a frame of N samples $(x_0, \ldots, x_{N-1})$:

$$a[k] \triangleq \frac{\sum_{n=k}^{N-1} x_n x_{n-k}}{\sum_{n=0}^{N-1} x_n^2} \tag{51}$$

The normalized autocorrelation has strong peaks for speech signals, which are robust against transient noises. We use the amplitude of the highest peak noted m_X as a feature

$$m_X \triangleq \max_k \{a[k]\} \tag{52}$$

where we disregard the initial peak of the self-correlation. Then, the feature is combined with the kurtosis to obtain the enhanced kurtosis f_X

$$f_X \triangleq m_X \cdot \log(1 + k_X). \tag{53}$$

The enhanced kurtosis is shown in Fig. 4, where the signal is an extract of the CENSREC-1-C dataset. One can observe that the enhanced kurtosis is more robust against transient noises compared to the kurtosis in the first five seconds (corresponding to walking steps in the background). One can also confirm that the desired behavior of the kurtosis for speech segments is not significantly altered by the combination with the normalized autocorrelation peak. The enhanced kurtosis is computed over overlapping windows of 256 samples at 8 kHz (32 ms) with an overlap of 50% (16 ms).

B. Online VB-EM-Based Classifier

For the VAD based on online EM (Section II), we make an assumption that each class (speech and noise) is locally distributed as Gaussian random variables. Instead of always assuming a model where "speech-and-noise" follow a bimodal distribution, we also estimate an additional model where only noise is assumed to be present. Both models are estimated with the same input data, using online VB-EM described in the previous section. At each frame, the online free energy of each model is compared, and the model with the highest free energy is used for classification. Thus, we design and implement a VAD method where both model comparison and classifier parameters are estimated online with the online VB-EM procedure. The overall scheme is depicted in Fig. 5, and summarized as follows.

1) The feature is computed frame per frame.
2) We conduct two online VB-EM in parallel, one with one component to model the situation "noise-only," and the other with two components, to model the situation "speech-and-noise." Both are updated frame per frame.
3) Compare the two models:
   a) $\hat{F}^p_{n,2} > \hat{F}^p_{n,1}$, we set the classifier to the "speech-and-noise" mode.
   b) When $\hat{F}^p_{n,2} < \hat{F}^p_{n,1}$, we set the classifier to the "noise-only" mode.
4) When the classifier is in the "speech-and-noise" mode, we apply the conventional Bayesian classifier with two classes, each modeled with a Gaussian, to the corresponding frame. The parameters of the Bayesian classifier are updated frame-by-frame through the VB-EM procedure corresponding to two Gaussians, considering all the data seen so far, with more weight on recent data through the learning parameter γ_n.
5) When the classifier is in "noise-only" mode, the signal is judged as noise only.

Both classifier parameters and classifier reliability are estimated from the same underlying statistical model. Although having several modes for the classifier has already been proposed for VAD, for example in [6], our proposed method uses the online free energy as a criterion for switching the modes instead of some heuristics.

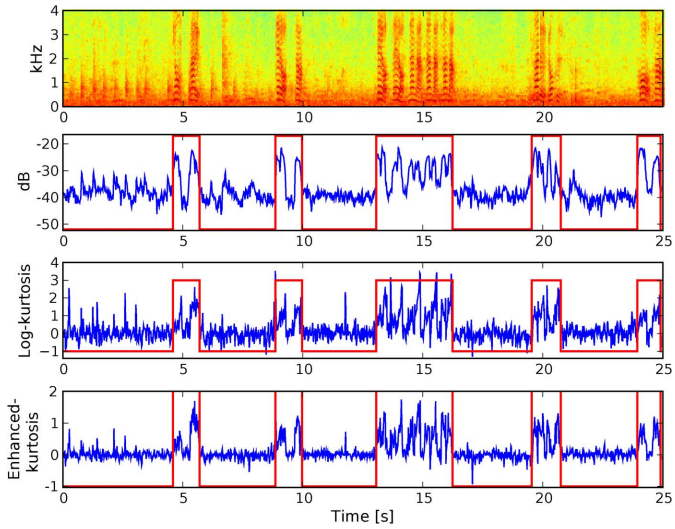

Fig. 4. Sample of speech from CENSREC-1 (high SNR). Spectrogram (top), energy (second), log-kurtosis (third) and proposed feature (fourth). This is the same signal used as in Fig. 1.

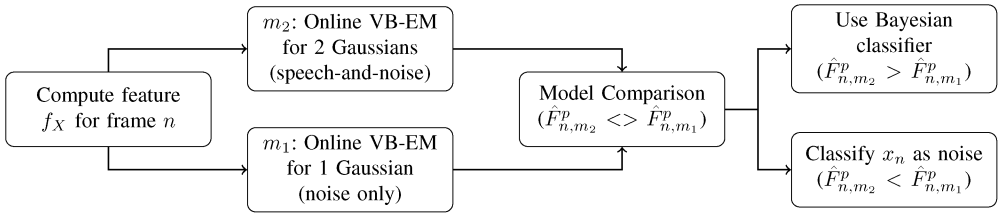

Fig. 5. Proposed scheme based on online VB-EM.

C. Practical Considerations

The VB-EM procedure requires initial values for the posterior, as well as prior hyper-parameter values.

- The posterior hyper-parameters are initialized using a conventional k-means method, as often used in conventional EM methods. The first two seconds of the signals are used as input data for the k-means procedure.
- We used a weak prior ($\tau_0 = 1$), identically set for each component. In this setting, it was observed that the hyper-parameter values did not have significant influence on the results.

After the initialization, the latency of the classification is one frame. It was also observed that not updating $\hat{\theta}_n$ at the beginning of the online procedure led to a more stable behavior. In our implementation, the value of $\hat{\theta}_n$ is kept unchanged for the first 60 frames, i.e., approximately one second. The implemented method uses a learning rate suggested in [14]

$$\gamma_n \triangleq \frac{1}{\sum_{s=1}^{n} \left(\prod_{t=s+1}^{n} \delta_t\right)} \tag{54}$$

with δ_t defined as

$$1 - \delta_t \triangleq \frac{1}{(t-2)k + t_0}. \tag{55}$$

As discussed in [14] (Section III-C), t_0 controls the learning speed in the early stage, and k controls the effective decreasing mode in the later stage. The value for the learning parameter γ_n was found to be more sensitive to the value t_0. Small t_0 values imply that old data are forgotten more quickly in the online procedure, and may prevent a stable behavior. In our implementation, t_0 was set to 100 and k to 0.01 (as in [14]).

Compared to online EM, the main additional cost of the online VB-EM procedure is the computation of $\overline{\theta}(\hat{\tau}_n, \hat{\alpha}_n)$, which needs to be done at every frame in real time. However, the cost is not prohibitive since online VB-EM is applied to very simple models in our case. In our current implementation which was mostly done with the Python interpreted language and not carefully optimized, the whole VAD process takes a real time factor (RTF) equal to 0.03.[5]

V. Evaluation

For the evaluation purpose, we use the CENSREC-1-C database [26], a Japanese dataset specifically designed for VAD evaluations. This database consists of noisy contiguous digit utterances in Japanese. The audio signals were recorded in two kinds of noisy environments (street and restaurant), in both low and high SNR conditions. For each of these conditions, close recordings (where the speakers were using a headset microphone) and remote recordings (where the speakers were approximately 50 cm away from the microphone) are available [26]. The speech signals were recorded with natural noise, i.e., noises were not artificially mixed. In the case of restaurant noise, low-SNR data were recorded in crowded situations, whereas high-SNR data were recorded in less busy hours. In the case of street noise, low-SNR data were recorded near an actual highway, whereas high-SNR data were recorded further away. For each recording situation (remote versus close recordings), approximately two hours of data are available. Each (noise type, noise level) combination makes data of approximately 30 minutes.

The results are measured by two kinds of error rates: false alarm rate (FAR: ratio of noise frames detected as speech divided by the number of noise frames) and false rejection rate (FRR: ratio of speech frames detected as noise divided by the number of speech frames).

To evaluate the effectiveness of the proposed method, we compare it against other published methods for VAD:

1) The first method corresponds to the method in [8]: it uses the same enhanced feature, and online VB-EM for classification without model comparison,[6] i.e., only one model is estimated. It is equivalent to the proposed method except that it assumes that the model with two components is always selected. It is used to confirm whether the proposed model comparison scheme actually improves the classification performance.
2) The second method is the one proposed by Sohn *et al.* [27], [28]. This method is based on modeling silence/speech state transitions using an HMM, with a noise model estimated on the first frames of the signal, which are assumed to be noise only. The features are frequency-band energies (the underlying statistical model for the speech-and-noise feature is similar to the one proposed in [29]).

A. Evaluation on Close Recordings

We first compare the results in Table I for the close recording data. For each method, the threshold was set so as to get roughly equal error rates (FAR = FRR) on the whole data set (all SNRs and noise types). Compared with online VB-EM without model comparison, an overall improvement is observed with the proposed method: both FAR and FRR are reduced to a more than half. It was observed that online model comparison helps at the

[5]Comparable or not slower than the Sohn's method used for evaluation in the next section.

[6]The method in [8] is based on online EM instead of online VB-EM. We preliminary confirmed that both online EM and online VB-EM without model comparison performed comparatively.

TABLE I
RESULTS OF THE ONLINE VB-EM WITH MODEL COMPARISON VAD ON THE **Close** DATASET, COMPARED TO THE SOHN VAD AND ONLINE VB-EM WITHOUT MODEL COMPARISON, BY SNR LEVELS

Online VB-EM with model comparison (proposed)	High SNR	Low SNR	Average
FAR	4.2 %	4.1 %	4.2 %
FRR	3.8 %	5.5 %	4.7 %
Online VB-EM without model comparison	High SNR	Low SNR	Average
FAR	8.5 %	10.4 %	9.5 %
FRR	10.8 %	11.1 %	11.0 %
Sohn VAD	High SNR	Low SNR	Average
FAR	3.9 %	3.9 %	3.9 %
FRR	1.9 %	6.8 %	4.1 %

TABLE II
RESULTS OF THE ONLINE VB-EM WITH MODEL COMPARISON VAD ON THE **Close** DATASET, COMPARED TO THE SOHN VAD, BY NOISE TYPE

Online VB-EM with model comparison (proposed)	High SNR	Low SNR	Average
FAR	4.0 %	4.3 %	4.2 %
FRR	4.6 %	4.7 %	4.7 %
Sohn VAD	Restaurant	Street	Average
FAR	7.1 %	0.7 %	3.9 %
FRR	4.2 %	4.5 %	4.4 %

TABLE III
RESULTS OF THE ONLINE VB-EM WITH MODEL COMPARISON VAD ON THE **Remote** DATASET, COMPARED TO THE SOHN VAD, AND ONLINE VB-EM WITHOUT MODEL COMPARISON, BY SNR TYPE

Online VB-EM with model comparison (proposed)	High SNR	Low SNR	Average
FAR	17.2 %	21.4 %	19.3 %
FRR	8.6 %	29.6 %	19.2 %
Online VB-EM without model comparison	High SNR	Low SNR	Average
FAR	15.2 %	26.8 %	21.0 %
FRR	13.1 %	30.9 %	22.0 %
Sohn VAD	High SNR	Low SNR	Average
FAR	19.9 %	31.1 %	25.5 %
FRR	16.0 %	33.3 %	24.7 %

beginning of each file, when no speech data are available. It prevents spurious detection of speech if only noise data are available, but without requiring the usually made assumption that only noise is input at first.

We display in Table II the comparison by the noise type (restaurant and street combining low and high SNRs). Both proposed method and the Sohn VAD show comparable results, but the proposed method performs more consistently across different conditions (in particular FAR for restaurant noise versus street noise).

B. Evaluation on Remote Recordings

Next, we tested on the remote recording data. The results are first given in terms of low and high SNRs in Table III, and in terms of restaurant and street noise in Table IV. The tendency in the close recordings is observed in the remote recordings as well, and the proposed method significantly outperforms the Sohn VAD. In particular, the sensitivity to the noise type is more significant for the Sohn VAD, whereas the proposed method is more consistent against different noise conditions.

TABLE IV
RESULTS OF THE PROPOSED VAD ON THE **Remote** DATASET, COMPARED TO THE STATISTICAL VAD, BY NOISE TYPE

Online VB-EM with model comparison (proposed)	High SNR	Low SNR	Average
FAR	24.6 %	14.9 %	19.7 %
FRR	17.6 %	20.6 %	19.1 %
Sohn VAD	Restaurant	Street	Average
FAR	49.1 %	1.6 %	25.4 %
FRR	14.3 %	33.8 %	24.1 %

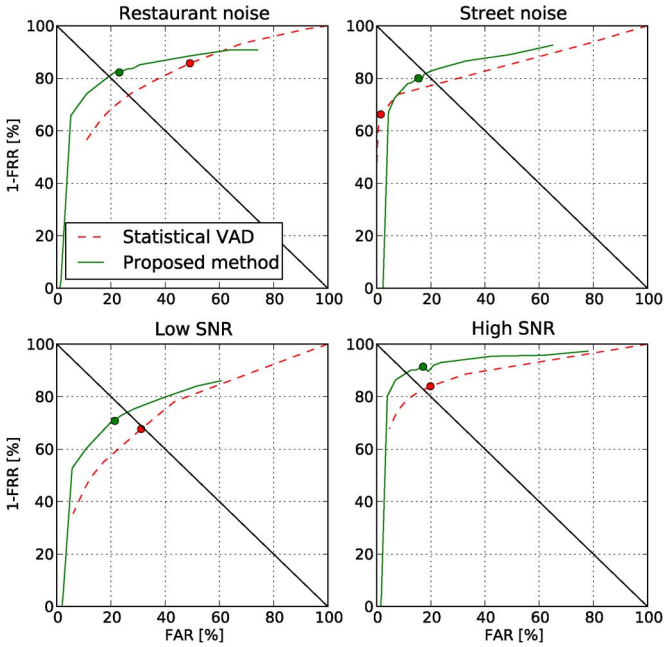

Fig. 6. ROC curves of the proposed method (solid line) against the Sohn VAD (dashed line). The plain circle represents the operating point corresponding to overall equal error rates.

This behavior is confirmed on receiver operating characteristic (ROC) curves displayed in Fig. 6. It is observed that the performance of the proposed method is consistently better than that of the Sohn VAD in all conditions, except for the high-SNR street-noise case, where both methods are comparable. It is also shown that the operating point corresponding to the equal error rates, displayed by a plain circle, is consistent for the four cases in the proposed method, suggesting robustness of the method.

VI. CONCLUSION

An enhanced scheme to online, unsupervised VAD based on online VB-EM is proposed. It considers two parallel statistical models for classification, one for noise-only and the other for speech-and-noise. Both models are estimated in an online manner. The online free energy, an online approximation of the log-evidence in the VB framework, is used to assess the classifier's reliability, and to decide which model to be used at a given time frame. The decision level is adapted automatically from the data, without the need for an *a priori* knowledge of the noise level. Thus, we can avoid some heuristics and complexity made in conventional VAD methods in an unsupervised context. The performance of the proposed method has been evaluated on the CENSREC-1-C database, and the proposed method improved the standard online EM and outperformed the other VAD methods.

APPENDIX I
ONLINE EM FOR MIXTURE OF GAUSSIANS

For the particular case of a mixture of K Gaussians, the latent data h_n may be defined as a discrete random variable such as $p(h_n = k)$ is the probability for the n^{th} observation to be in component k. Noting $w = w_1, \ldots, w_K$ the weights, $\mu = \mu_1, \ldots, \mu_K$ the means, and $\Omega = \Omega_1, \ldots, \Omega_K$ the precisions (inverse of variance) of the normal components, the complete data model is defined as

$$p(x_n, h_n; w, \mu, \Omega) \triangleq \sum_{k=1}^{K} \delta_{h_n,k} w_k \mathcal{N}(x_n; \mu_k, \Omega_k) \tag{56}$$

where $\delta_{i,j}$ is the Kronecker delta, and $\mathcal{N}(.; \mu_k, \Omega_k)$ the normal density of mean μ_k and precision matrix Ω_k

$$\mathcal{N}(x_n; \mu_k, \Omega_k) \triangleq \sqrt{\frac{|\Omega_k|}{(2\pi)^d}} \exp\left\{-\frac{1}{2}(x_n - \mu_k)'\Omega_k(x_n - \mu_k)\right\} \tag{57}$$

where d is the dimension of the random variable x_n, $|.|$ is the determinant of a square matrix, and x'_n the transpose of vector x_n. Rewriting the complete data model to follow a form similar to (5)

$$\begin{aligned}\ln p(x_n, h_n; w, \mu, \Omega) = &\sum_{k=1}^{K} \delta_{h_n,k}\left(\ln(w_k) + \frac{\ln|\Omega_k| - \mu'_k\Omega_k\mu_k}{2}\right) \\ &+ \delta_{h_n,k}\left(\mu'_k\Omega_k x_n - \frac{x'_n\Omega_k x_n}{2}\right) + \text{const.}\end{aligned} \tag{58}$$

From (58), noting $\theta = \theta_1, \ldots, \theta_K$, with θ_k defined as

$$\theta_k \triangleq \begin{pmatrix} \ln w_k + \frac{1}{2}\left(\ln|\Omega_k| - \mu'_k\Omega_k\mu_k\right) \\ \mu'_k\Omega_k \\ -\frac{\Omega_k}{2} \end{pmatrix} \tag{59}$$

and the corresponding sufficient statistics $s = (s_1, \ldots, s_K)$, with s_k defined as follows:

$$s_k(x_n, h_n) = \begin{pmatrix} s_{k,1}(x_n, h_n) \\ s_{k,2}(x_n, h_n) \\ s_{k,3}(x_n, h_n) \end{pmatrix} = \begin{pmatrix} \delta_{k,h_n} \\ \delta_{k,h_n} x_n \\ \delta_{k,h_n} x_n x'_n \end{pmatrix} \tag{60}$$

we obtain a parametrization of a GMM as a EHF defined in (5). The derivation of the conventional EM and online EM follows this parametrization. Noting $\overline{w}_k(x_n; \theta^i)$ the responsibilities:

$$\overline{w}_k(x_n; \theta^i) \triangleq p(h_n = k | x_n; \theta^i) \tag{61}$$

the per-component averaged SS $\overline{s}_k(x_n; \theta^i)$ for a GMM are

$$\overline{s}_k(x_n; \theta^i) = \begin{pmatrix} \overline{s}_{k,1}(x_n, \theta^i) \\ \overline{s}_{k,2}(x_n, \theta^i) \\ \overline{s}_{k,3}(x_n, \theta^i) \end{pmatrix} = \begin{pmatrix} \overline{w}_k(x_n; \theta^i) \\ \overline{w}_k(x_n; \theta^i) x_n \\ \overline{w}_k(x_n; \theta^i) x_n x'_n \end{pmatrix}. \tag{62}$$

Following the convention of (8), we note

$$\begin{aligned}\overline{s}_{k,1}(x; \theta^i) &\triangleq \overline{w}_k(x; \theta^{i+1}) \triangleq \frac{1}{N}\sum_n \overline{w}_k(x_n; \theta^i) \\ \overline{s}_{k,2}(x; \theta^i) &\triangleq \frac{1}{N}\sum_n \overline{w}_k(x_n; \theta^i) x_n \\ \overline{s}_{k,3}(x; \theta^i) &\triangleq \frac{1}{N}\sum_n \overline{w}_k(x_n; \theta^i) x_n x'_n.\end{aligned} \tag{63}$$

The usual update equations for M-step [(11)] applied to a GMM may be written as follows for iteration i:

$$\forall k \in 1, \ldots, K, \quad \begin{cases} w_k^i = \overline{s}_{k,1}(x; \theta^i) \\ \mu_k^i = \frac{\overline{s}_{k,2}(x;\theta^i)}{\overline{s}_{k,1}(x;\theta^i)} \\ \Omega_k^i = \frac{\overline{s}_{k,3}(x;\theta^i)}{\overline{s}_{k,1}(x;\theta^i)} - \mu_k^i {\mu_k^i}' \end{cases} \tag{64}$$

Eq. (64) gives an explicit formulation for the function f defined in (10), and the averaged SS can be computed from (62). The online EM update equations for the GMM follow directly. For a GMM, (14) becomes

$$\begin{aligned}\hat{s}_{n+1,k} &= \begin{pmatrix} \hat{s}_{n+1,k,1} \\ \hat{s}_{n+1,k,2} \\ \hat{s}_{n+1,k,3} \end{pmatrix} \\ &= \hat{s}_{n,k} \\ &\quad + \gamma_{n+1}\left(\begin{pmatrix} \bar{w}_k(x_{n+1}; \hat{\theta}_n) \\ \bar{w}_k(x_{n+1}; \hat{\theta}_n) x_{n+1} \\ \bar{w}_k(x_{n+1}; \hat{\theta}_n) x_{n+1} x'_{n+1} \end{pmatrix} - \hat{s}_{n,k}\right)\end{aligned} \tag{65}$$

and the online estimation of the parameters is derived from (64) by replacing the averaged SS $\bar{s}$ with the approximated SS $\hat{s}$ as defined by (65)

$$\forall k \in 1, \ldots, K, \quad \begin{cases} \hat{w}_{n+1,k} = \hat{s}_{n+1,k,1} \\ \hat{\mu}_{n+1,k} = \frac{\hat{s}_{n+1,k,2}}{\hat{s}_{n+1,k,1}} \\ \hat{\Omega}_{n+1,k} = \frac{\hat{s}_{n+1,k,3}}{\hat{s}_{n+1,k,1}} - \hat{\mu}_{n+1,k}\hat{\mu}'_{n+1,k} \end{cases} \tag{66}$$

In summary, the online EM equations for the GMM can be retrieved from the conventional EM formulas once they are written in the form of (14). The approximated SS $\hat{s}_n$ [(65)] can be derived directly from the conventional averaged SS $\bar{s}$ used for the batch EM [(62)], and the online estimates of the GMM parameters are obtained from replacing the averaged SS $\bar{s}$ by the approximated SS $\hat{s}_n$ in the M step [20].

APPENDIX II
ONLINE VB-EM FOR MIXTURE OF GAUSSIANS

For a mixture of K Gaussians, a conjugate prior is the product of a Dirichlet prior on the weights and K Normal–Wishart priors on the mean and precision (μ_k, Ω_k) for each component. The prior is written as follows:

$$\begin{aligned}&p_0(\theta; \lambda_0, m_0, \beta_0, a_0, B_0) \\ &\quad = \mathcal{D}(w_1, \ldots, w_K; \lambda_{0,1} + 1, \ldots, \lambda_{0,K} + 1) \\ &\qquad \times \prod_k \mathcal{NW}(\mu_k, \Omega_k; m_{0,k}, \beta_{0,k}, a_{0,k}, B_{0,k})\end{aligned} \tag{67}$$

where we keep the conventions of Appendix I, i.e., $\lambda_0 = \lambda_{0,1}, \ldots, \lambda_{0,K}$, and similar notations for m_0, β_0, a_0 and B_0. $\lambda_{0,k}$, $\beta_{0,k}$, and $a_{0,k}$ are strictly positive scalars, $m_{0,k}$ a vector of dimension d, and $B_{0,k}$ is a positive definite matrix of dimension d. The Dirichlet prior $\mathcal{D}$ is defined as

$$\mathcal{D}(w_1, \ldots, w_K; \lambda_0 + 1) \triangleq \frac{\Gamma\left(\sum_k \lambda_{0,k} + K\right)}{\prod_k \Gamma(\lambda_{0,k} + 1)} \prod_k w_k^{\lambda_{0,k}} \tag{68}$$

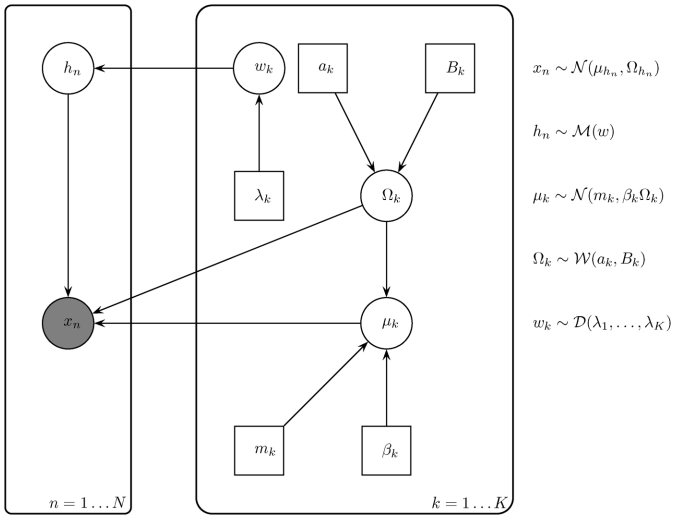

Fig. 7. Graphical model representing the Bayesian mixture of Gaussians model. Square nodes correspond to deterministic variables (hyper-parameters), circles are random variables and dashed nodes are the observed values.

with Γ the Gamma function. The Normal–Wishart $\mathcal{NW}$ is defined as

$$\begin{aligned}\mathcal{NW}(\mu_k, \Omega_k; m_{0,k}, \beta_{0,k}, a_{0,k}, B_{0,k}) \\ \triangleq \mathcal{N}(\mu_k; m_{0,k}, \beta_{0,k}\Omega_k)\mathcal{W}(\Omega_k; a_{0,k}, B_{0,k})\end{aligned} \quad (69)$$

with the Wishart defined as [23]

$$\begin{aligned}\mathcal{W}(\Omega_k; a_{0,k}, B_{0,k}) &\triangleq C_k |\Omega_k|^{\frac{a_{0,k}-d-1}{2}} e^{-\frac{1}{2}\operatorname{tr} B_{0,k}\Omega_k} \\ C_k^{-1} &\triangleq B_{0,k}^{-a_{0,k}/2} 2^{da_{0,k}/2} \pi^{d(d-1)/4} \\ &\quad \cdot \prod_{j=1}^{d} \Gamma\left(\frac{a_{0,k}+1-j}{2}\right).\end{aligned} \quad (70)$$

The model is summarized in Fig. 7.

The prior p_0 can be rewritten in the same form as (31) with the following parametrization [14]:

$$\begin{aligned}\tau_0 &\triangleq \sum_k \lambda_{0,k} \\ \alpha_{0,k,1} &\triangleq \lambda_{0,k} = (a_{0,k} - d) = \beta_{0,k} \\ \alpha_{0,k,2} &\triangleq (a_{0,k} - d) m_{0,k} \\ \alpha_{0,k,3} &\triangleq (a_{0,k} - d) m_{0,k} m_{0,k}' + B_{0,k}.\end{aligned} \quad (71)$$

Those equations define the natural parametrization for the conjugate prior hyper-parameters, as for one component, (31) becomes

$$\begin{aligned}\ln p_0(\theta; \alpha_0) &\propto \left\langle \theta, \begin{pmatrix}\alpha_{0,2} \\ \alpha_{0,3}\end{pmatrix} \right\rangle - \alpha_{0,1}\psi(\theta) \\ &= \left\langle \begin{pmatrix}\mu\Omega \\ -\frac{1}{2}\Omega\end{pmatrix}, \begin{pmatrix}(a_0 - d)m_0 \\ (a_0 - d)m_0 m_0' + B_0\end{pmatrix} \right\rangle \\ &\quad - \frac{\alpha_{0,1}}{2}\left(\mu'\Omega\mu - \ln|\Omega|\right) \\ &= \frac{a_0 - d}{2}(\mu - m_0)'\Omega(\mu - m_0) + \frac{\ln|\Omega|}{2} \\ &\quad + d\frac{a_0 - d - 1}{2}\ln|\Omega| - \frac{1}{2}\operatorname{tr} B_0\Omega \\ &\propto \ln \mathcal{NW}(\mu, \Omega; m_0, a_0 - d, a_0, B_0)\end{aligned} \quad (72)$$

where we dropped the component index k for notational purpose. This gives the explicit formulas for the M step through (39) and (40), since the posterior has the same form as the prior. The expression for $\bar{s}(x; \overline{\theta}^{i+1})$ is exactly the same as for the conventional EM algorithm in the MLE framework [(63)], but parametrized by the averaged parameters $\overline{\theta}^{i+1}$. For mixtures of Gaussians, noting λ_k^i, m_k^i, β_k^i, a_k^i, and B_k^i the hyper-parameters of the posterior after the i^{th} iteration, the explicit form of $\overline{\theta}^{i+1}$ is found from the responsibilities $q_{h_n}^i(h_n)$ [23], [30]

$$\begin{aligned}q_{h_n}^i(h_n = k) &\propto \tilde{w_k}^i \tilde{\beta_k}^i \\ &\quad \times \exp\left\{-\frac{d}{2\beta_k^i} - \frac{\beta_k^i}{2}\left(x_n - m_k^i\right)' B_k^i \left(x_n - m_k^i\right)\right\}\end{aligned} \quad (73)$$

where

$$\begin{aligned}\ln \tilde{w_k}^i &\triangleq F\left(\lambda_k^i\right) - F\left(\sum_{k'=1}^{K} \lambda_{k'}^i\right) \\ \ln \tilde{\beta_k}^i &\triangleq \sum_{j=1}^{d} F\left(\frac{\lambda_k^i + 1 - j}{2}\right) - d\ln 2 - \ln\left|B_k^i\right|\end{aligned} \quad (74)$$

and $F \triangleq \Gamma'/\Gamma$ is the digamma function. Equation (73) and (74) give the explicit formulas for the E-step of the VB-EM procedure. Equation (71) with (39) and (40) gives the explicit formulas for the M-step

$$\begin{aligned}\lambda_k^{i+1} &= \lambda_{0,k} + N_k \\ \beta_k^{i+1} &= \beta_{0,k} + N_k \\ a_k^{i+1} &= a_{0,k} + N_k \\ m_k^{i+1} &= \frac{\beta_{0,k} m_{0,k}}{\beta_{0,k} + N_k} + \frac{N_k \bar{x}_k}{\beta_{0,k} + N_k} \\ B_k^{i+1} &= B_{0,k} + N_k S_k \\ &\quad + \frac{\beta_0 N_k}{\beta_0 + N_k}(\bar{x}_k - m_{0,k})(\bar{x}_k - m_{0,k})'\end{aligned} \quad (75)$$

where

$$\begin{aligned}N_k &\triangleq N\bar{s}_{k,1}(x; \overline{\theta}^{i+1}) \\ \bar{x}_k &\triangleq \frac{\bar{s}_{k,2}(x; \overline{\theta}^{i+1})}{\bar{s}_{k,1}(x; \overline{\theta}^{i+1})} \\ S_k &\triangleq \frac{1}{N_k}\sum_n \overline{w}_k(x_n; \overline{\theta}^{i+1})(x_n - \bar{x}_k)(x_n - \bar{x}_k)' \\ &= \frac{\bar{s}_{k,3}(x; \overline{\theta}^{i+1})}{\bar{s}_{k,1}(x; \overline{\theta}^{i+1})} - \bar{x}_k \bar{x}_k'.\end{aligned} \quad (76)$$

Appendix III
Variational Bayes as Parametrized Free Energy Optimization

We note q_h^p and q_θ^p the parametrized posterior

$$\begin{aligned}q_h^p(h; \eta) &\triangleq p(h|x; \eta) \\ \ln q_\theta^p(\theta; \nu, \beta) &\triangleq \langle \theta, \beta \rangle - \nu\psi(\theta) - \zeta(\nu, \beta)\end{aligned}$$

where $p(h|x;\eta)$ is the responsibilities under parameters η and q_θ^p has the same functional form as the prior p_0, (31). q_h^p and q_θ^p are by definition equal to q_h^i and q_θ^i at the end of the i^{th} E step and M step iteration respectively, when $\eta = \overline{\theta}^{i+1}$, $\nu = \tau_0 + N$, and $\beta = \alpha_0 + \bar{s}(x;\overline{\theta}^{i+1})$. The parametrized free energy is then defined as the free energy F_m [(28)] with $\tilde{q}$ replaced by $q_h^p \cdot q_\theta^p$

$$\begin{aligned} F_m^p(\eta,\nu,\beta) &\triangleq F_m\left[q_h^p \cdot q_\theta^p\right] \\ &= \mathbb{E}_{q_h^p \cdot q_\theta^p}\left[\ln p(x,h,\theta)\right] - \mathbb{E}_{q_h^p}\left[\ln q_h^p(h;\eta)\right] \\ &\quad - \mathbb{E}_{q_\theta^p}\left[\ln q_\theta^p(\theta;\nu,\beta)\right]. \end{aligned} \tag{77}$$

Whereas the free energy F_m was a functional of argument $\tilde{q}$, the parametrized free energy F_m^p is a function of parameters (η,ν,β). This parametrized free energy F_m^p can be rewritten as follows:

$$\begin{aligned} F_m^p(\eta,\nu,\beta) = \mathbb{E}_{q_h^p \cdot q_\theta^p}\left[\ln \frac{p(x,h|\theta)}{p(x,h;\eta)}\right] + \ln p(x;\eta) \\ - \mathbb{E}_{q_\theta^p}\left[\ln \frac{q_\theta^p(\theta;\nu,\beta)}{p_0(\theta;\tau_0,\alpha_0)}\right]. \end{aligned} \tag{78}$$

The derivation of F_m^p w.r.t. each of its arguments can be done explicitly [14]. Ignoring the terms which do not depend on η

$$\begin{aligned} F_m^p(\eta,\nu,\beta) \propto \ln p(x;\eta) + N\left\langle \bar{s}(x;\eta), \bar{\theta}(\nu,\beta) - \eta \right\rangle \\ + N\Psi(\eta) + \text{const.} \end{aligned} \tag{79}$$

and using the following property of the score function for EF

$$\frac{\partial \ln p(x;\eta)}{\partial \eta} = N\bar{s}(x;\eta) - N\Psi(\eta) \tag{80}$$

one obtains [14]

$$\frac{\partial F_m^p}{\partial \eta}(\eta,\nu,\beta) = N\frac{\partial \bar{s}(x;\eta)}{\partial \eta}.\left(\bar{\theta}(\nu,\beta) - \eta\right). \tag{81}$$

Similarly, ignoring the terms which do not depend on (ν,β) in (78)

$$\begin{aligned} F_m^p(\eta,\nu,\beta) &\propto \left\langle N\bar{s}(x;\eta) + \alpha_0 - \beta, \bar{\theta}(\nu,\beta)\right\rangle \\ &\quad + (\nu - \tau_0 - N)\overline{\Psi}(\nu,\beta) + \zeta(\nu,\beta) + \text{const.} \\ &\propto \left\langle N\bar{s}(x;\eta) + \alpha_0 - \beta, \frac{\partial \zeta}{\partial \beta}(\nu,\beta)\right\rangle \\ &\quad - (\nu - \tau_0 - N)\frac{\partial \zeta}{\partial \nu}(\nu,\beta) + \zeta(\nu,\beta) \\ &\quad + \text{const.} \end{aligned} \tag{82}$$

where we note

$$\overline{\Psi}(\nu,\beta) \triangleq \mathbb{E}_{q_\theta^p}\left[\Psi(\theta)\right] = -\frac{\partial \zeta}{\partial \nu}(\nu,\beta) \tag{83}$$

and we used (41)

$$\bar{\theta}(\nu,\beta) = \frac{\partial \zeta}{\partial \beta}(\nu,\beta). \tag{84}$$

Hence, one obtains [14]

$$\begin{aligned} \frac{\partial F_m^p}{\partial \nu}(\eta,\nu,\beta) = \left\langle N\bar{s}(s;\eta) + \alpha_0 - \beta, \frac{\partial^2 \zeta}{\partial \nu \partial \beta}(\nu,\beta)\right\rangle \\ -(\nu - \tau_0 - N)\frac{\partial^2 \zeta}{\partial \nu^2}(\nu,\beta) \end{aligned} \tag{85}$$

and

$$\begin{aligned} \frac{\partial F_m^p}{\partial \beta}(\eta,\nu,\beta) = \left\langle N\bar{s}(s;\eta) + \alpha_0 - \beta, \frac{\partial^2 \zeta}{\partial^2 \beta}(\nu,\beta)\right\rangle \\ -(\nu - \tau_0 - N)\frac{\partial^2 \zeta}{\partial \beta \partial \nu}(\nu,\beta). \end{aligned} \tag{86}$$

The derivation of F_m^p w.r.t. each argument can be succinctly written as

$$\frac{\partial F_m^p}{\partial \eta}(\eta,\nu,\beta) = N\frac{\partial \bar{s}(x;\eta)}{\partial \eta}.\left(\bar{\theta}(\nu,\beta) - \eta\right) \tag{87}$$

$$\begin{pmatrix} \frac{\partial F_m^p}{\partial \nu}(\eta,\nu,\beta) \\ \frac{\partial F_m^p}{\partial \beta}(\eta,\nu,\beta) \end{pmatrix} = H_\zeta(\nu,\beta).\begin{pmatrix} \bar{s}(x;\eta) + \alpha_0 - \beta \\ N + \tau_0 - \nu \end{pmatrix} \tag{88}$$

where H_ζ is the Hessian of ζ

$$H_\zeta(\nu,\beta) \triangleq \begin{pmatrix} \frac{\partial^2 \zeta}{\partial \beta^2}(\nu,\beta) & \frac{\partial^2 \zeta}{\partial \beta \partial \nu}(\nu,\beta) \\ \frac{\partial \partial^2 \zeta}{\partial \partial \nu \partial \beta}(\nu,\beta) & \frac{\partial^2 \zeta}{\partial \alpha^2}(\nu,\beta) \end{pmatrix}. \tag{89}$$

Thus, the VB-EM version of (33)–(35) are derived from the stationary points of F_m^p [14]. Maximizing F_m^p w.r.t. each of its arguments is equivalent to the VB-EM procedure for complete data models in the EF. The E-step is derived from (87) and the M-step is derived from (88). If the E-step $\eta = \bar{\theta}(\tau,\alpha)$ is incorporated in the parametrized free energy, this can be rewritten as [14]

$$\begin{aligned} \frac{\partial F_m^p}{\partial \nu}\left(\eta = \bar{\theta}(\tau,\alpha),\tau,\alpha\right) = 0 \\ \frac{\partial F_m^p}{\partial \beta}\left(\eta = \bar{\theta}(\tau,\alpha),\tau,\alpha\right) = 0. \end{aligned} \tag{90}$$

ACKNOWLEDGMENT

The authors would like to thank Dr. K. Ishizuka and Dr. M. Fujimoto for useful discussions, comments and the reference implementation of the methods described in [9] and [27]. The authors would also like to thank Dr. M. Sato for useful discussions on online VB-EM.

REFERENCES

[1] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[2] B. Kingsbury, G. Saon, L. Mangu, M. Padmanabhan, and R. Sarikaya, "Robust speech recognition in noisy environments: The 2001 IBM Spine evaluation system," in *Proc. ICASSP*, 2002, pp. I-53–I-56.
[3] J. K. Shah, A. N. Iyer, B. Y. Smolenski, and R. E. Yantorno, "Robust voiced- unvoiced classification using novel features and Gaussian mixture model," in *Proc. ICASSP*, 2004.
[4] S. Basu, "A linked-HMM model for robust voicing and speech detection," in *Proc. ICASSP*, 2003, pp. 816–819.
[5] E. Dong, G. Liu, Y. Zhou, and X. Zhang, "Applying support vector machine to voice activity detection," in *Proc. 6th Int. Conf. Signal Process. (ICSP)*, 2002, pp. 1124–1127.
[6] K. Li, M. S. S. Swamy, and M. O. Ahmad, "An improved voice activity detection using high order statistics," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 965–974, Sep. 2005.
[7] I. Shafran and R. Rose, "Robust speech detection and segmentation for real-time ASR applications," in *Proc. ICASSP*, 2003, pp. 432–435.
[8] D. Cournapeau and T. Kawahara, "Evaluation of real-time voice activity detection based on high order statistics," in *Proc. Interspeech*, 2007.
[9] E. Nemer, R. Goubran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the LPC residual domain," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 3, pp. 217–231, Mar. 2001.

[10] M. J. Beal and Z. Ghahramani, "The variational Bayesian EM algorithm for incomplete data: With application to scoring graphical model structures," *Bayesian Statist.*, vol. 7, pp. 453–464, 2002.
[11] D. MacKay, *Information Theory, Inference and Learning Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2003.
[12] S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda, "Variational Bayesian estimation and clustering for speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 4, pp. 365–381, Jul. 2004.
[13] H. J. Kushner and G. G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, 2nd ed. Berlin, Germany: Springer-Verlag, 2003.
[14] M. Sato, "Online model selection based on the variational Bayes," *Neural Comput.*, vol. 13, pp. 1649–1681, 2001.
[15] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc.: Ser. B (Statisti. Methodol.)*, vol. 39, no. 1, pp. 1–38, 1977.
[16] M. Sato and S. Ishii, "On-line EM algorithm for the normalized Gaussian network," *Neural Comput.*, vol. 12, pp. 407–432, 2000.
[17] O. Cappé, M. Charbit, and E. Moulines, "Recursive EM algorithm with applications to DOA estimation," in *Proc. ICASSP*, 2006, pp. 664–667.
[18] C. F. J. Wu, "On the convergence properties of the EM algorithm," *Ann. Statist.*, vol. 11, pp. 95–103, 1983.
[19] D. R. Cox, *Principles of Statistical Inference*. Cambridge, U.K.: Cambridge Univ. Press, 2006.
[20] O. Cappé and E. Moulines, "Online EM algorithm for latent data models," *J. R. Statist. Soc.: Ser. B (Statist. Methodol.)*, vol. 73, no. 3, pp. 593–613, 2009.
[21] H. Attias, "Inferring parameters and structure of latent variable models by variational Bayes," in *Proc. 15th Conf. Uncertainty Artif. Intell.*, 1999, pp. 21–30 [Online]. Available: http://citeseer.ist.psu.edu/attias99inferring.html
[22] D. J. C. Mackay, "Bayesian interpolation," *Neural Comput.*, vol. 4, pp. 415–447, 1992.
[23] C. Bishop, *Pattern Recognition and Machine Learning*. Berlin, Germany: Springer-Verlag, 2006.
[24] M. Beal, "Variational algorithms for approximate Bayesian inference," Ph.D. dissertation, Gatsby Computational Neuroscience Unit, Univ. College London, London, U.K., 2003.
[25] D. Cournapeau and T. Kawahara, "Using variational Bayes free energy for unsupervised voice activity-detection," in *Proc. IEEE-ICASSP*, 2008.
[26] N. Kitaoka, T. Yamada, S. Tsuge, C. Miyajima, T. Nishiura, M. Nakayama, Y. Denda, M. Fujimoto, K. Yamamoto, T. Takiguchi, and S. Nakamura, "CENSREC-1-C: Development of evaluation framework for voice activity detection under noisy environment," (in Japanese) IPSJ SIG SLP, 2006, Tech. Rep.
[27] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett*, vol. 6, no. , pp. 1–3, Jan. 1999.
[28] J. Ramirez, J. Segura, C. Benitez, A. de la Torre, and A. Rubio, "A new Kullback-Leibler vad for speech recognition in noise," *IEEE Signal Process. Lett.*, vol. 11, no. 2, pp. 266–269, Feb. 2004.
[29] Y. Ephraim and D. Malah, "Speech enhancement using minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Audio, Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.
[30] T. Minka, "Using lower bounds to approximate integrals," Carnegie Mellon Univ., Pittsburgh, PA, 2001, Tech. Rep.

David Cournapeau received the M.Sc. degree from Universite Pierre Marie Curie (UPMC), Paris, France, in 2003, specializing in acoustics and music signal processing, the M.E. degree from Telecom Paris Tech in 2004, and the Ph.D. degree from Kyoto University, Kyoto, Japan, in 2009, under the supervision of Prof. Kawahara.

His research interests include digital signal processing, speech and music signal processing, signal representation, statistical signal processing, and pattern recognition.

Shinji Watanabe (M'03) received the B.S., M.S., and Dr.Eng. degrees from Waseda University, Tokyo, Japan, in 1999, 2001, and 2006, respectively.

In 2001, he joined Nippon Telegraph and Telephone Corporation (NTT) and has since been working at NTT Communication Science Laboratories, Kyoto, Japan. From January to March in 2009, he was a Visiting Scholar at the Georgia Institute of Technology, Atlanta, in Dr. Juang's laboratory. His research interests include Bayesian learning, pattern recognition, and speech and spoken language processing.

Dr. Watanabe is a member of the Acoustical Society of Japan (ASJ) and the Institute of Electronics, Information, and Communications Engineers (IEICE). He received the Awaya Award from the ASJ in 2003, the Paper Award from the IEICE in 2004, the Itakura Award from ASJ in 2006, and the TELECOM System Technology Award from the Telecommunications Advancement Foundation in 2006.

Atsushi Nakamura received the B.E., M.E., and Dr.Eng. degrees from Kyushu University, Fukuoka, Japan, in 1985, 1987 and 2001, respectively.

In 1987, he joined Nippon Telegraph and Telephone Corporation (NTT), where he engaged in the research and development of network service platforms, including studies on application of speech processing technologies into network services, at Musashino Electrical Communication Laboratories, Tokyo, Japan. From 1994 to 2000, he was with Advanced Telecommunications Research (ATR) Institute, Kyoto, Japan, as a Senior Researcher, working on the research of spontaneous speech recognition, construction of spoken language database, and development of speech translation systems. Since April, 2000, he has been with NTT Communication Science Laboratories, Kyoto, Japan. His research interests include acoustic modeling of speech, speech recognition and synthesis, spoken language processing systems, speech production and perception, computational phonetics and phonology, and application of learning theories to signal analysis and modeling.

Dr. Nakamura is a member of the Machine Learning for Signal Processing (MLSP) Technical Committee, as well as served as a Vice Chair of the Signal Processing Society Kansai Chapter. He is also a member of the Institute of Electronics, Information and Communication Engineering (IEICE) and the Acoustical Society of Japan (ASJ). He received the IEICE Paper Award in 2004, and received twice the Telecom-Technology Award of The Telecommunications Advancement Foundation, in 2006 and 2009.

Tatsuya Kawahara (M'91–SM'08) received the B.E., M.E., and Ph.D. degrees in information science from Kyoto University, Kyoto, Japan, in 1987, 1989, and 1995, respectively.

In 1990, he became a Research Associate in the Department of Information Science, Kyoto University. From 1995 to 1996, he was a Visiting Researcher at Bell Laboratories, Murray Hill, NJ. Currently, he is a Professor in the Academic Center for Computing and Media Studies and an Adjunct Professor in the School of Informatics, Kyoto University. He is also an Invited Researcher at ATR Spoken Language Communication Research Laboratories. He has published more than 150 technical papers covering speech recognition, spoken language processing, and spoken dialogue systems. He has been managing several speech-related projects in Japan including a free large-vocabulary continuous speech recognition software project (http://julius.sourceforge.jp/).

Dr. Kawahara received the 1997 Awaya Memorial Award from the Acoustical Society of Japan and the 2000 Sakai Memorial Award from the Information Processing Society of Japan. From 2003 to 2006, he was a member of the IEEE SPS Speech Technical Committee. He was a general Co-Chair of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU-2007).